第 14 卷 第 3 期 集 成 技 术 Vol. 14 No. 3
2025 年 5 月 JOURNAL OF INTEGRATION TECHNOLOGY May 2025

引文格式：
窦铭扬, 耿艳娟, 杨佳彬. 基于聚焦注意力机制的对齐回归手部姿态估计网络 [J]. 集成技术, 2025, 14(3): 64-77.
Dou MY, Geng YJ, YANG JB. Alignment regression hand pose estimation network based on focused attention mechanism [J]. Journal of Integration Technology, 2025, 14(3): 64-77.

基于聚焦注意力机制的对齐回归手部姿态估计网络

窦铭扬[1,2] 耿艳娟[1*] 杨佳彬[1,3]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](中国科学院大学 北京 100049)
[3](南方科技大学 深圳 518055)

摘 要 基于 RGB 图像的手部姿态估计在动态手势识别和人机交互领域有着广泛的应用前景。然而，现有方法面临诸多挑战，如手部自相似性程度高和关键点分布极为密集等问题，导致难以以较低的计算成本实现高精度的预测，进而导致在复杂场景中的表现存在局限性。鉴于此，本文提出一种基于 YOLOv8 网络的二维手部姿态估计模型 —— FAR-HandNet。该模型巧妙地融合了聚焦线性注意力模块、关键点对齐策略和回归残差拟合模块，有效地增强了对小目标区域 (如手部) 的特征捕捉能力，同时减少了自相似性对手部关键点定位精度的不良影响。此外，回归残差拟合模块基于流生成模型对关键点残差分布进行拟合，极大地提升了回归模型的精度。本文实验在卡内基梅隆大学全景数据集 (CMU) 和 FreiHAND 数据集上展开。实验结果表明，FAR-HandNet 在参数量和计算效率方面优势明显，与现有方法相比，在不同阈值下的正确关键点概率表现优异。同时，该模型的推理时间仅需 32 ms。消融实验进一步证实了各模块的有效性，充分验证了 FAR-HandNet 在手部姿态估计任务中的有效性和优越性。

关键词 手部姿态估计；注意力机制；回归网络
中图分类号 TP399 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20241030001
CSTR: 32239.14.j.issn.2095-3135.20241030001

Alignment Regression Hand Pose Estimation Network Based on Focused Attention Mechanism

DOU Mingyang[1,2] GENG Yanjuan[1*] YANG Jiabin[1,3]

[1](*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
[3](*Southern University of Science*, *Shenzhen* 518055, *China*)
[*]*Corresponding Author*: *yj.geng@siat.ac.cn*

Abstract Hand pose estimation based on RGB images holds wide application prospects in dynamic gesture

收稿日期：2024-10-30 **修回日期**：2025-03-11
基金项目：国家自然科学基金项目 (62373345)；深圳市医学研究专项项目 (D2402013)；深圳市基础研究重点项目 (JCYJ20220818101602005)
作者简介：窦铭扬，硕士研究生，研究方向为手部姿态估计、人机交互； 耿艳娟 (通讯作者)，博士，副研究员，博士研究生导师，研究方向为生物医学信号处理、神经机器接口与人机交互、运动感觉功能康复技术与机理等，E-mail：yj.geng@siat.ac.cn； 杨佳彬，硕士研究生，研究方向为肌电信号处理。

recognition and human-computer interaction. However, existing methods face challenges such as high hand self-similarity and densely distributed keypoints, making it difficult to achieve high-precision predictions with low computational costs, thereby limiting their performance in complex scenarios. To address these challenges, this paper proposes a 2D hand pose estimation model named FAR-HandNet, based on the YOLOv8 network. The model ingeniously integrates a focused linear attention module, a keypoint alignment strategy, and a regression residual fitting module, effectively enhancing feature capture capabilities for small target regions (e.g., hands) while mitigating the adverse effects of self-similarity on the localization accuracy of hand keypoints. Additionally, the regression residual fitting module leverages a flow-based generative model to fit the residual distribution of keypoints, significantly improving regression precision. Experiments were conducted on the Carnegie Mellon University panorama dataset (CMU) and the FreiHAND dataset. Results demonstrate that FAR-HandNet exhibits remarkable advantages in parameter size and computational efficiency. Compared to existing methods, it achieves superior performance in the percentage of correct keypoints under varying thresholds. Furthermore, the model achieves an inference time of only 32 ms. Ablation studies further validate the effectiveness of each module, conclusively verifying the efficacy and superiority of FAR-HandNet in hand pose estimation tasks.

Keywords　hand pose estimation; attention mechanism; regression network

Funding　This work is supported by National Natural Science Foundation of China (62373345), Shenzhen Medical Research Special Project (D2402013), and Shenzhen Governmental Basic Research Grant (JCYJ20220818101602005)

1　引　言

在当今数字化时代，人机交互技术蓬勃发展，手部姿态估计作为其中一项关键技术，为智能交互体验提供了有力支撑。随着科技的不断进步，低成本且便携式的图像采集设备越来越普及，手部姿态估计技术顺势在多个领域开启了广泛应用。回顾过去数年，深度图像数据集的引入为手部姿态估计领域带来了显著进步。然而，深度传感器自身存在短板，其分辨率与有效范围相对受限，并且对环境光照条件极为“挑剔”[1]，通常仅能在室内环境稳定发挥作用，这无疑严重束缚了手部姿态估计技术在各个领域的发展。近年来，卷积神经网络得到广泛关注，基于RGB 图像的手部姿态估计方法也随之发展壮大，与之相伴的是 RGB 图像数据集数量的节节攀升，单目 RGB 方法备受瞩目[2]。从成本考量，RGB 相机亲民的价格优势明显；从应用广度来看，它摆脱了深度传感器的诸多限制，室内外皆宜。但与此同时，基于 RGB 图像的手部姿态估计也面临着各种挑战。手部自身结构特殊，如具有很强的自相似性、关节分布密集和姿态复杂等特性导致现有方法在复杂环境下难以精准“命中”关键点，进而导致难以实现精确预测。

当下，主流的手部姿态估计方法主要分为基于热图和基于回归两种。基于热图的方法的核心思路是预测各关键点的大致方位，为每个关键点在所有可能位置生成对应的概率热度图[3-4]，进而选取概率峰值所在位置作为关键点的最终定位。例如：Newell 等[5]提出一种堆叠沙漏网络架构，每个沙漏模块最终会输出一组热图。Wei 等[6]提出的卷积姿态机 (convolutional pose machines，

CPM) 为各关键点量身打造一组热图，确保每个关键点在自身对应的热图中脱颖而出，拥有位置概率最大值；Wang 等[7]提出一种两阶段级联卷积神经网络架构，首阶段专注于手部掩码的预测，次阶段巧妙基于输出特征图和首阶段生成的手部掩码，对手关节位置进行估计，并在该阶段运用基于 CPM 架构的热图回归策略；Wang 等[8]提出的 SRHandNet 方法借助编码器-解码器架构开展热图回归，精准获取手部位置的边界框坐标和手部关键点位置，同时利用同一网络进行周期推理，提升手部姿态估计精度。然而，基于热图的方法并非尽善尽美，取概率最大值这一操作使其无法采用端到端的训练方法；加之深度神经网络降采样的“副作用”，热图分辨率与输入图片相比大打折扣，引发不可逆的量化误差，关节位置精度受限。若试图通过提高热度图分辨率来改善，则又会陷入内存和计算开销剧增的困境。

基于回归的方法则另辟蹊径，应用端到端框架，全力学习从输入图像到关键点位置的直接映射关系。如此一来，既无须费心维护高分辨率热图，大大减轻了内存和计算负担，又能直接输出坐标值，巧妙避开量化误差。Toshev 等[9]将 2D 人体姿态估计问题进行巧妙转化，将传统的图像处理与模板匹配难题转变为依托卷积神经网络的图像特征提取和关键点坐标回归新问题，运用回归范式对被遮挡的人体关节点进行精准估计。这一方法不仅省却了热图生成步骤，还保留了对全局约束的深刻理解能力。Geng 等[10]更是洞察到精确回归关键点坐标的关键所在——需特别聚焦关键点周围区域，进而提出结构式关键点回归这一直接坐标回归方法。该方法采用自适应的卷积结构，潜心学习关键点周围像素的特征，并通过多分支结构设计，让每个分支针对特定关键点，利用自适应卷积深挖关键点周围的像素特征，实现对关键点位置的精准回归。多阶段回归策略的运用可进一步提高关键点坐标的精度，让多阶段直接回归方法焕发新活力。Carreira 等[11]提出的自我修正模型同样独具匠心，利用输入到输出的联合空间深入学习特征提取器对联合空间中蕴藏的丰富结构化信息进行精细建模。该模型引入自顶向下的反馈机制，凭借反馈错误预测逐步修正初始解，在迭代错误反馈 (iterative error feedback，IEF) 过程中不断提高模型精度。然而，回归方法虽然在时间效率上表现出色，但面对复杂手势时，准确性较差，而且容易忽视关键点之间的内在关联信息，精度弱于基于热图的方法。

为攻克回归方法精度欠佳以及未能充分考量关键点间联系的难题，本文在 YOLOv8 的主干网络 CSPDarkNet 中将 C2f 替换为聚焦线性注意力模块 (focusing linear attention module，FLAM)。通常情况下，人手在图像中所占比例较小，关键点分布相对集中。本文对主干网络的优化不仅降低了计算成本，还能高效获取小目标区域的特征信息，更能精准提取关键特征。与此同时，本文提出关键点对齐策略 (key points alignment，KPA)，将关键点的空间关系和类别信息融入回归框架中，筛选出候选关键点。这一策略不仅考虑了关键点之间的紧密联系，还极大地降低了手部自相似性对手部关键点定位准确性的负面影响，并显著加快模型训练进程。此外，本文还基于残差似然估计重构回归残差拟合模块 (regression residual fitting module，RrFM)，引入流生成模型拟合潜在的输出分布，从而生成关键点的概率分布。如此一来，无须费力维护热图，便能比肩基于热图方法的性能，大幅提升模型的预测精度。

综上所述，本文提出一种单阶段端到端的 2D 手部姿态估计方法，基于聚焦注意力机制的对齐回归手部姿态估计网络 (FAR-HandNet)，该方法能敏锐捕捉小目标区域，通过融合关键点间的空间和类别信息，以及对回归范式的创新性重

构，提升模型的预测精度，有望为手部姿态估计领域提供新的解决方案。

2　基于聚焦注意力机制的对齐回归手部姿态估计网络

2.1　整体架构

FAR-HandNet 包括聚焦线性注意力、关键点对齐和回归残差拟合等 3 个模块，如图 1 所示。在主干网络中将 C2f 替换为 FLAM，使模型更注重小目标区域的特征细节 (如指尖)，提高模型的表达能力，并降低计算量；KPA 基于各关键点的空间关系和类别信息，极大地降低了手部自相似性对关键点定位准确性的影响，并加快了模型的训练进程；RrFM 通过利用流生成模型拟合回归误差与基础分布之间的残差概率分布，可提高回归模型对关键点预测的准确性。通过拟合残差分布，可使回归模型达到热图模型的效果。图 1 中，$G_\phi(\bar{x})$ 为通过流生成模型学习的分布，$Q(\bar{x})$ 为标准正态分布。

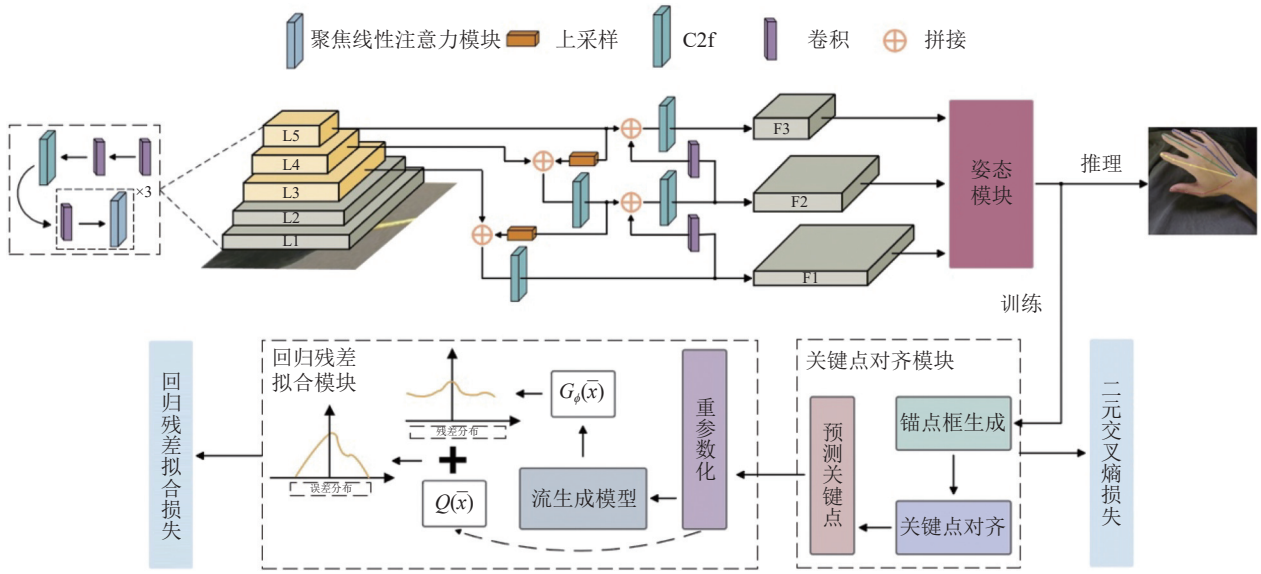


图 1　FAR-HandNet 整体网络架构

Fig. 1　FAR-HandNet overall network architecture

2.2　特征提取器的优化

2.2.1 基于聚焦线性注意力的主干网络优化设计

在 YOLOv8 的主干网络 CSPDarkNet 中，本文将传统的 C2f 模块替换为 FLAM，以提高网络对小目标 (如手部) 的特征提取能力，同时降低计算量。改进后的 CSPDarkNet-VT 如图 2 所示。

同时，FLAM 的引入在以下几个方面对网络产生了显著影响。用 FLAM 替换原先的 C2f 模块，使输入的特征图经过 FLAM 的处理后，能以更高效的方式捕捉重要的局部特征，尤其是小目标区域 (如手指尖) 的细节。这一改变使得网络能更聚焦于高维特征，减少对背景噪声的关注，从而提升对小目标的识别精度。

在 FLAM 的处理下，网络能有效降低自注意力机制中传统 Softmax 的计算复杂度，从而加速训练和推理过程。FLAM 利用线性注意力机制的 $O(Nd^2)$ 复杂度替代传统注意力机制的 $O(N^2d)$ 复杂度，同时通过聚焦算法增强对重要区域的关注力，使网络在提取特征时，能更准确地捕捉手部的关键特征。基于深度卷积 (depthwise convolution，

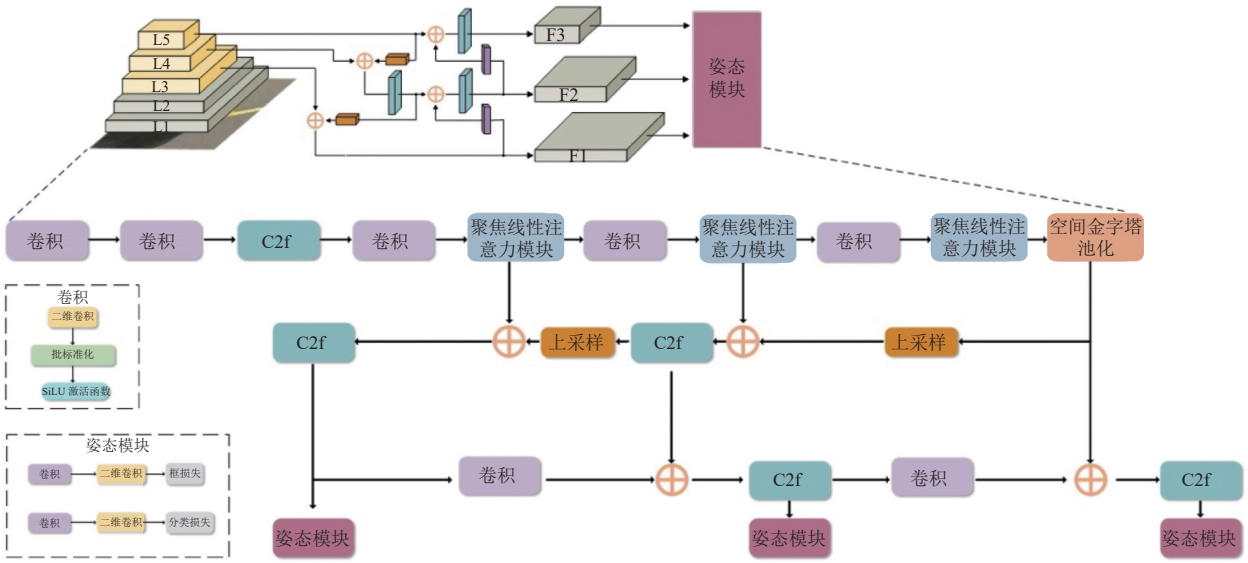


图 2 CSPDarkNet-VT 的整体网络模型

Fig. 2 Overall network model of CSPDarkNet-VT

DWC) 和逐点卷积 (pointwise convolution，PWC) 模块，FLAM 进一步增强了特征多样性，确保了空间和通道特征的丰富性，从而提高了网络的表达能力。

由于 FLAM 能高效提取局部和通道之间的特征，因此通过 FLAM 优化的网络输出更注重小目标区域的精细信息。这使得 FLAM 在预测阶段能提高关键点的定位精度，尤其是在处理复杂的手部姿态时。与传统的 C2f 模块相比，FLAM 对手部自相似性和复杂姿态的适应性更强，可提高经过优化的特征提取器在手部关键点检测任务中的表现。

对 CSPDarkNet 进行优化，不仅可有效降低计算量，还可显著提升特征提取的准确性和细节捕捉能力，在处理小目标时，表现尤其突出。

2.2.2 聚焦线性注意力模块

近年来，Transformer 模型在计算机视觉领域取得了显著进展，在复杂任务中表现尤其出色。然而，图像数据的高维特性导致全局自注意力计算复杂度为 $O(N^2d)$，在处理高分辨率图像时会造成极大的计算负担[12]。因此，本文在 FAR-HandNet 中采用了一种基于聚焦算法的线性注意力机制，如图 3 所示。该机制通过简化矩阵运算顺序可有效将计算复杂度降低至 $O(Nd^2)$，同时保留了对重要特征区域的高效捕捉能力，在处理小目标 (如手部) 时表现尤其出色[13-14]。

(1) 线性注意力

在视觉 Transformer 中，自注意力表示如下：

$$\boldsymbol{Q}=X\boldsymbol{W}_{\boldsymbol{Q}},\boldsymbol{K}=X\boldsymbol{W}_{\boldsymbol{K}},\boldsymbol{V}=X\boldsymbol{W}_{\boldsymbol{V}}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i=\sum_{j=1}^{N}\frac{\text{Sim}(\boldsymbol{Q}_i,\boldsymbol{K}_j)}{\sum_{j=1}^{N}\text{Sim}(\boldsymbol{Q}_i,\boldsymbol{K}_j)}\boldsymbol{V}_j \tag{1}$$

其中，每个自注意力头给定 N 个输入 $X\in\mathbb{R}^{N\times C}$；投影矩阵 $\boldsymbol{W}_{\boldsymbol{Q}},\boldsymbol{W}_{\boldsymbol{K}},c\in\mathbb{R}^{C\times C}$，$\boldsymbol{Q}$ 为当前需要关注的目标位置或信息需求，用于主动检索其他位置的关联性；$\boldsymbol{K}$ 存储序列中每个位置的标识信息，用于匹配 $\boldsymbol{Q}$ 的查询需求；$\boldsymbol{V}$ 包含每个位置的实际信息内容，最终通过权重加权后输出；$\text{Sim}(\cdot,\cdot)$ 为相似函数，在 Softmax 注意力中的相似函数为 $\text{Sim}(\boldsymbol{Q},\boldsymbol{K})=\exp(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d})$。线性注意力机制通过改变矩阵计算顺序降低计算量，表示如下：

$$\text{Sim}(\boldsymbol{Q}_i,\boldsymbol{K}_j)=\varnothing(\boldsymbol{Q}_i)\varnothing(\boldsymbol{K}_j)^{\mathrm{T}} \tag{2}$$

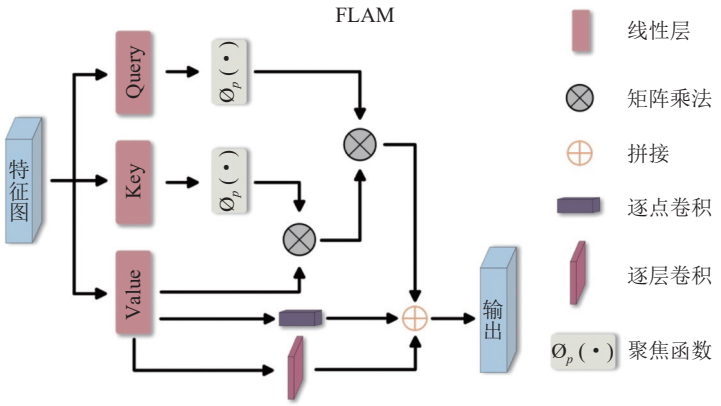


图 3　FLAM 网络框架

Fig. 3　FLAM networking framework

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})_i=\frac{\varnothing(\boldsymbol{Q}_i)\left(\sum_{j=1}^{N}\varnothing(\boldsymbol{K}_j)^{\mathrm{T}}\boldsymbol{V}_j\right)}{\varnothing(\boldsymbol{Q}_i)\left(\sum_{j=1}^{N}\varnothing(\boldsymbol{K}_j)^{\mathrm{T}}\right)} \tag{3}$$

(2) 聚焦算法

Softmax 注意力实际上提供了一种非线性重新加权机制，使得传统注意力机制的注意力矩阵在某些区域 (如前景物体) 的分布特别尖锐。相比之下，线性注意力的分布更平滑，不能关注信息更丰富的区域。因此提出聚焦算法，使相似的查询密钥对更接近，不相似的查询密钥对相距更远。聚焦算法映射函数 f_p 表示如下：

$$\text{Sim}(\boldsymbol{Q}_i,\boldsymbol{K}_j)=\varnothing_p(\boldsymbol{Q}_i)\varnothing_p(\boldsymbol{K}_j)^{\mathrm{T}} \tag{4}$$

$$\varnothing_p(x)=f_p(\text{ReLU}(x)),\ f_p(x)=\frac{\|x\|}{\|x^{**p}\|}x^{**p} \tag{5}$$

其中，x^{**p} 表示逐元素求 x 的 p 次幂。为保证式 (5) 中输入的非负性和分母的有效性，在映射函数 f_p 前使用 ReLU 函数。

(3) 特征多样性

在 DeiT-Tiny 所使用的 Transformer 层中 tokens 数量 $N=14\times14$，且注意力矩阵具有满秩 (即 196)[15]。然而线性注意力的注意力矩阵受 tokens 的数量 N 和每个头的通道维度 d 的限制，很难做到满秩。

$$\begin{aligned}\text{rank}\left(\varnothing(\boldsymbol{Q})\varnothing(\boldsymbol{K})^{\mathrm{T}}\right)&\leqslant\min\{\text{rank}(\varnothing(\boldsymbol{Q})),\text{rank}(\varnothing(\boldsymbol{K}))\}\\&\leqslant\min\{N,d\}\end{aligned} \tag{6}$$

在常见视觉 Transformer 中，d 通常小于 N，注意力矩阵秩的上限被限制，表明线性注意力的注意力矩阵有多个行存在严重的同质化，这将导致聚合特征之间存在相似性。为解决此类问题，将 DWC 添加到注意力矩阵中：

$$\text{Attention}=\varnothing(\boldsymbol{Q})\varnothing(\boldsymbol{K})^{\mathrm{T}}\boldsymbol{V}+\text{DWC}(\boldsymbol{V}) \tag{7}$$

DWC 能关注空间上的局部特征，即使在线性注意力中两个查询向量相同，也能从不同的局部特征获得不同的输出，保证了特征的多样性。但该模块仅能获得空间上的局部特征，忽略了通道间的特征信息，从而会导致特性多样性不够完善，本文尝试加入 PWC 模块解决此问题：

$$\text{Attention}=\varnothing(\boldsymbol{Q})\varnothing(\boldsymbol{K})^{\mathrm{T}}\boldsymbol{V}+\text{DWC}(\boldsymbol{V})+\text{PWC}(\boldsymbol{V}) \tag{8}$$

其中，PWC 能提取通道间的特征信息，丰富特征多样性[16]。

(4) 聚焦线性注意力机制

综上所述，本文实现了一种带有特征聚焦能力的线性注意力机制，并在特征多样性方面加以改进，更新后的模块可表示如下：

$$\begin{aligned}\text{Attention}&=\text{Sim}(\boldsymbol{Q},\boldsymbol{K})\boldsymbol{V}\\&=\varnothing_p(\boldsymbol{Q})\varnothing_p(\boldsymbol{K})^{\mathrm{T}}\boldsymbol{V}+\text{DWC}(\boldsymbol{V})+\text{PWC}(\boldsymbol{V})\end{aligned} \tag{9}$$

此模块通过改变矩阵计算顺序将计算复杂度从 $O(N^2d)$ 降至 $O(Nd^2)$，并加入聚焦算法使该注意力的表达能力可以与 Softmax 相媲美，巧妙地将 DWC 与 PWC 联合使用，保证了该模块的特征多样性。

2.3 关键点对齐

为应对手部自相似性对关键点定位带来的干扰，本文提出了关键点对齐策略。该策略通过将关键点的空间布局信息与类别信息结合，筛选出高质量的候选关键点，极大地减少了错误定位的概率。此外，关键点对齐策略有效利用了锚点和正负样本的动态匹配机制，加速了模型训练，并提高了预测精度，在密集关键点的手势中表现尤其突出。

2.3.1 无锚框策略

在之前有锚框的工作中，针对不同数据集需要手动设置锚框的长宽比和锚框数量等属性，采用的匹配机制导致极端尺度相对于适中尺度被匹配到的频率更低，使得网络的泛化性降低[17]。有锚框策略为了兼顾多尺度下的预测能力，推理得到的预测框也相对较多，在输出处理时，非极大值抑制计算也会更耗时。

YOLOv8 中采用无锚框的策略，训练中直接学习各种框的形状，推理时根据学习到的边框距离与锚点中心位置拟合物体尺寸[18]。无锚框策略通过不依赖数据集中的先验知识，使网络对“物体形状”有更好的表达能力，泛化能力更强。因此无锚框策略在运动物体和尺寸不一的检测上精度有所提升，同时检测被遮挡物体时也更灵活。

2.3.2 关键点对齐策略

在姿态模块之前将会输出 3 个不同尺度的特征图，共产生 8 400 个锚框，通过姿态模块进行关键点预测，每个锚框都会预测 21 个关键点，因此本节通过设计关键点对齐策略筛选最佳的预测关键点。通常情况下，真实关键点附近区域内的点均可视为关键点。在处理不同输入图像中的目标物体时，由于目标物体 (如手部) 的大小并不相同，且关键点的分布存在差异，真实关键点包围框需要动态调整。因此本文提出一种基于目标物体关键点间距离的动态包围框设计方法。具体而言，包围框的边长根据以下几个关键点类型及其相互距离计算得出。对于每个关键点 i(i 从 0 到 $n-1$，其中 n 为关键点总数)，根据其类型和相对位置计算关键点之间的距离 D_i，表示如下：

$$\begin{cases} D_i=\sqrt{(x_{\mathrm{gt}}^{i}-x_{\mathrm{gt}}^{i-1})^2+(y_{\mathrm{gt}}^{i}-y_{\mathrm{gt}}^{i-1})^2}, & \text{if } i \bmod 4=0 \text{ and } i\neq 0 \\ D_i=\sqrt{(x_{\mathrm{gt}}^{i+1}-x_{\mathrm{gt}}^{i})^2+(y_{\mathrm{gt}}^{i+1}-y_{\mathrm{gt}}^{i})^2}, & \text{if } i \bmod 4=1 \text{ or } i=0 \\ D_i=\dfrac{1}{2}\sqrt{(x_{\mathrm{gt}}^{i+1}-x_{\mathrm{gt}}^{i})^2+(y_{\mathrm{gt}}^{i+1}-y_{\mathrm{gt}}^{i})^2}+\dfrac{1}{2}\sqrt{(x_{\mathrm{gt}}^{i}-x_{\mathrm{gt}}^{i-1})^2+(y_{\mathrm{gt}}^{i}-y_{\mathrm{gt}}^{i-1})^2}, & \text{otherwise} \end{cases} \tag{10}$$

其中，$(x_{\mathrm{gt}}^{i},y_{\mathrm{gt}}^{i})$ 为输入图像中第 i 个关键点的坐标，gt 为真实关键点，选择 0.3 倍 D_i 作为第 i 个真实关键点包围框的边长，目的是最大程度地减弱框之间的重叠。同样地，选择第 i 个真实关键点包围框的边长为第 i 个预测关键点创建包围框。计算预测关键点包围框和真实关键点包围框之间的交并比 (intersection over union，IOU) 值。

此外，受正负样本动态策略的启发，关键点对齐策略不仅需考虑预测关键点与真实关键点之间的空间关联，还需考虑二者的类别相似度 S：

$$S(P_i,\mu)=\mathrm{e}^{-\left(\frac{(P_i-\mu)^2}{2\sigma^2}\right)}+\epsilon \tag{11}$$

其中，P_i 为预测关键点的类别得分；μ 为真实关键点的标注类别；σ 为缩放程度，取 $\sigma=0.1$；ϵ 为极小值，取 $\epsilon=10^{-6}$，为避免相似度为 0。P_i 与 μ 越接近，$S(P_i,\mu)$ 越接近 1，反之接近 0。

同时考虑关键点的类别得分和预测关键点包围框与真实关键点包围框之间的 IOU 值，可更好地表征关键点的对齐程度[19]：

$$t=S^{\alpha}\times \mathrm{IOU}^{\beta} \tag{12}$$

其中，α 和 β 用来控制 S 和 IOU 对对齐度量的影响。t 将引导模型从关键点对齐的角度动态关注高质量的锚点。关键点对齐策略通过获得前 k 个 t 筛选最佳的预测关键点，一般情况下，$k=10$。

2.3.3 二元交叉熵损失函数

二元交叉熵损失衡量了模型预测的类别与真实标签之间的差异，即用于衡量模型预测的准确程度。因此本文使用二元交叉熵损失度量预测关键点与真实关键点之间的差异，表示如下：

$$\mathcal{L}_{\text{bce}}=\frac{1}{n}\sum_{i=1}^{n}-[y_i\log(\text{sigmoid}(x_i))+(1-y_i)\log(1-\text{sigmoid}(x_i))] \tag{13}$$

其中，n 为关键点个数；y_i 为真实关键点类别标签；x_i 为预测关键点类别得分。

2.4　回归残差拟合模块

在回归方法中，由于更接近真实分布的密度函数会带来更好的回归性能[20]，因此本文尝试使用一种带有残差对数似然估计的回归残差拟合模块 RrFM 捕获潜在的输出分布。RrFM 通过流生成模型学习分布变化，并结合重新参数化设计优化回归效果。

2.4.1 标准回归范式

标准回归范式是将 L1 或 L2 损失应用于不同任务的回归输出 μ。在基于回归的姿态估计任务中，μ 通常为预测关键点的坐标 (X,Y)。在给定输入图像 I 中，回归模型将预测真实标注出现在位置 x 的概率分布 $P_\theta(x/I)$，其中，θ 为回归模型的参数；μ_{g} 为输入图像中的真实标注。从最大似然估计 (maximum likelihood estimation，MLE) 的角度分析，模型的学习目标是不断优化参数 θ，使 $P_\theta(x/I)|_{x=\mu_g}$ 最大，因此该过程的损失函数可表示如下：

$$\mathcal{L}_{\text{mle}}=-\log P_\theta(x/I)|_{x=\mu_g} \tag{14}$$

假设概率分布为高斯分布，通过回归模型预测的 μ 和 σ 构造概率密度函数：$P_\theta(x/I)=\frac{1}{\sqrt{2\pi}\sigma}\text{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$，使真实标签出现在预测位置上的可能性最大化。损失函数表示如下：

$$\mathcal{L}=-\log P_\theta(x/I)|_{x=\mu_g}\propto\log\sigma+-\frac{(\mu_g-\mu)^2}{2\sigma^2} \tag{15}$$

当假定该分布的方差 σ 为常量时，该损失函数为 L2 损失：$\mathcal{L}=(\mu_g-\mu)^2$。若假设的概率分布为拉普拉斯分布，且方差 σ 为常量，则损失函数退化为 L1 损失。由此可知，损失函数与概率分布的形状相关，但仅通过上述的损失函数并不能得到准确的概率分布，因此提高概率密度函数的准确性能帮助模型更好地回归。

2.4.2 基于流生成模型的回归方法

流生成模型通过构造一个可逆映射和转换简单分布构造复杂分布。首先通过回归模型预测的 μ 和 σ 构造初始概率密度函数 $P_\theta(z/I)$。假定一个分布为高斯分布，其中 z 为随机变量，则 $P_\theta(z/I)=\frac{1}{\sqrt{2\pi}\sigma}\text{e}^{-\frac{(z-\mu)^2}{2\sigma^2}}$。通过流生成模型构造一个平滑且可逆映射的 $f_\varphi:\mathbb{R}^2\to\mathbb{R}^2$，将 z 转换为 x，即 $x=f_\varphi(z)$，其中 φ 为流生成模型中的可学习参数。通过该映射函数，变量 x 将服从 $P_{\theta,\varphi}(x/I)$，该分布取决于回归参数 θ 与流生成模型 f_φ，表示如下：

$$\log P_{\theta,\varphi}(x/I)=\log P_\theta(\text{z}/I)+\log\left|\det\frac{\partial f_\varphi^{-1}}{\partial x}\right| \tag{16}$$

其中，∂f_φ^{-1} 为 f_φ 的逆函数，即 $z=\partial f_\varphi^{-1}(x)$。$f_\varphi$ 在训练过程中不断学习，理论上可拟合任意分布，但均值 μ 与方差 σ 会根据输入的 I 而变化，因此利用流生成模型 f_φ 将零均值的初始分布 $\bar{z}\sim N(0,I)$ 映射到零均值的变形分布 $\bar{x}\sim P_\varphi(\bar{x})$，其中 $\bar{x}=(x-\mu)/\sigma$，μ 和 σ 通过回归模型 θ 预测，用于控制分布的位置和尺度。由此可知流生成模型 f_φ 对变形分布 $\bar{x}\sim P_\varphi(\bar{x})$ 的学习受 θ 和 φ 的影响，在训练初期，流生成模型估计的分布可能并不理想，这将影响目标分布。因此需要设计新的回归范式，通过流生成模型 f_φ 学习残差分布变

化，假设最优分布为 $P_{\mathrm{ed}}(\bar{x})$，则

$$\begin{aligned}\log P_{\mathrm{ed}}(\bar{x}) &= \log\left(\boldsymbol{Q}(\bar{x})\cdot\frac{P_{\mathrm{ed}}(\bar{x})}{s\cdot\boldsymbol{Q}(\bar{x})}\cdot s\right)\\ &= \log\boldsymbol{Q}(\bar{x})+\log\frac{P_{\mathrm{ed}}(\bar{x})}{s\cdot\boldsymbol{Q}(\bar{x})}+\log s\end{aligned}\tag{17}$$

其中，$\boldsymbol{Q}(\bar{x})=N(0,1)$；$\log\frac{P_{\mathrm{ed}}(\bar{x})}{s\cdot\boldsymbol{Q}(\bar{x})}$ 为残差似然项。为确保残差项保持分布形式，设置常量 s。本文假定 $\boldsymbol{Q}(\bar{x})$ 能大致拟合最优误差分布，但并不完美，而残差项是对此的补充。由于 $P_{\varphi}(\bar{x})$ 不断向 $P_{\mathrm{ed}}(\bar{x})$ 拟合，因此：

$$\log P_{\varphi}(\bar{x})=\log\boldsymbol{Q}(\bar{x})+\log G_{\varphi}(\bar{x})+\log s\tag{18}$$

其中，$G_{\varphi}(\bar{x})$ 为通过流生成模型学习到的分布；s 的值可用黎曼和近似，即 $s=\frac{1}{\int G_{\varphi}(\bar{x})\boldsymbol{Q}(\bar{x})\mathrm{d}x}$。此时 $G_{\varphi}(\bar{x})$ 不再学习成为整体的误差分布，而是尝试拟合为残差分布 $\frac{P_{\mathrm{ed}}(\bar{x})}{s\cdot\boldsymbol{Q}(\bar{x})}$，因此最终的总损失函数表示如下：

$$\begin{aligned}\mathcal{L}_{\mathrm{mle}} &= -\log P_{\theta,\varphi}(x/I)|_{x=\mu_g}\\ &= -\log P_{\varphi}(\overline{\mu_g})+\log\sigma\\ &= -\log\boldsymbol{Q}(\overline{\mu_g})-\log G_{\varphi}(\overline{\mu_g})-\log s+\log\sigma\end{aligned}\tag{19}$$

综上所述，RrFM 从最大似然估计的角度结合流生成模型提出了一种新颖且有效的回归范式，并通过拟合残差分布获得更精准的关键点估计。

3 实 验

3.1 数据集与实验环境

本文在卡耐基梅隆大学的全景手数据集 (CMU) 上评估所提出的模型[21]。该数据集有 14 817 个样本，对应于从全景工作室捕获的图像中人的右手。通过随机抽样将该数据集分为 8∶1∶1，分别对应训练集、验证集和测试集。为重复验证本文模型，本文还在 FreiHAND[22]数据集上评估了本文的手部姿态估计模型。该数据集包含 4×32 560 个训练集样本和 3 960 个测试样本，每个样本的尺寸为 224×224，并带有手部包围框和 21 个关键点的标注，还包括各种照明条件和背景。

实验硬件如下：Intel 酷睿 i9-14900K 处理器，配置两块 NVIDIA 系列的 4090 GPU 显卡、DDR5_64 GB 内存和 2 TB 固态硬盘。实验软件如下：操作系统为 Ubantu18.04，安装 Anaconda 和 PyTorch 深度学习框架，编辑语言为 Python。本文初始学习率设为 0.001，动量为 0.937，batch size 大小为 64 (每张卡分配 32 个样本)，总迭代次数为 300，并使用 Adam optimizer 进行优化。使用均值为 0 且方差为 0.01 的正常初始化器初始化所提出的架构的权重。

3.2 评估指标

本文利用正确关键点概率 (percentage of correct keypoints，PCK) 评估姿态估计模型的性能。PCK 为预测关键点与真实关键点之间的欧氏距离位于 σ 内的概率。D 个样本上第 i 个手关键点的 PCK 表示如下：

$$\mathrm{PCK}_T^i=\frac{1}{D}\sum_D\delta\left(\frac{d_i<T}{\max(w,h)}\right)\tag{20}$$

其中，d_i 为第 i 个预测关键点与真实值间的欧氏距离；$\delta(\cdot)$ 为指标函数；w 和 h 为包围框的长和宽；T 为误差阈值。

4 结果与讨论

首先，本文比较了 CSPDarkNet-VT 与主流特征提取器 (如 ResNet 和 EfficientNet)，实验结果表明：CSPDarkNet-VT 在参数量和计算效率上具有明显优势，在 FreiHAND 和 CMU 数据集上的小目标特征提取能力较优，为后续模型构建提供了支持；其次，本文评估了聚焦线性注意力机制的有效性，实验结果表明，其在降低计算成本的同时，能保持与传统 Softmax 注意力相近的精度，在小目标检测任务中表现更佳，进一步证明了其在手部姿态估计中的优越性；再次，本文将

FAR-HandNet 与多种手部姿态估计方法进行了对比实验，表明基于回归的方法在手部姿态估计中能达到，甚至超越热图方法的性能；最后，通过消融实验，本文分析了 FAR-HandNet 各模块的贡献，发现关键点对齐策略显著提升了模型的整体精度，同时各模块之间的协同作用也增强了模型的性能。

综上所述，本文通过多个实验验证了 FAR-HandNet 在手部姿态估计中的强大能力，为后续研究提供了重要的理论依据和实践参考。

4.1　特征提取器

在 2D 手部姿态估计的研究中，特征提取器的性能对模型的整体表现起关键作用。为全面评估 CSPDarkNet-VT 的性能，本文在 CMU 和 FreiHAND 这两个具有代表性的数据集上，将其与当前流行的 ResNet 和 EfficientNet 特征提取器进行了详细的比较分析。

在参数量方面，CSPDarkNet-VT 展现出显著优势。如表 1 所示，CSPDarkNet-VT 的参数量仅为 3.04 M，而 ResNet34 的参数量高达 23.79 M，EfficientNetB1 和 EfficientNetB2 的参数量分别为 8.42 M 和 9.55 M。参数量越低，模型对存储和计算资源的需求越低，这对在资源受限的设备上部署模型具有重要意义。在移动设备或嵌入式系统中，有限的内存和计算能力限制了大模型的应用，而 CSPDarkNet-VT 的低参数量特性使其能更好地适应这些环境。

计算效率是衡量特征提取器性能的另一个重要指标。从浮点运算数来看，CSPDarkNet-VT 同样表现出色，其浮点运算数为 8.7 G，远低于其他模型，表明 CSPDarkNet-VT 在处理图像时，所需的计算量更少，能更快地完成特征提取任务。在实时应用场景中，如实时人机交互，快计算速度是保证系统实时性的关键。

在验证聚焦线性注意力机制有效性的实验中，对传统 Softmax 注意力与聚焦线性注意力

表 1　多个最先进模型作为特征提取器在 CMU 数据集上的性能

Table 1　Performance of multiple state of the art models as feature extractors on the CMU dataset

模型	参数量/M	浮点运算数/G	PCK/%
ResNet34	23.79	61.3	86.34
EfficientNetB1	8.42	10.1	87.09
EfficientNetB2	9.55	11.5	90.25
CSPDarkNet	3.39	9.7	88.33
CSPDarkNet＋Soft-Attention	3.26	9.4	92.16
CSPDarkNet-VT (Our)	3.04	8.7	92.18

注：PCK 为 T＝0.1 时的值

在 CMU 和 FreiHAND 数据集上的性能分别进行了对比，结果如表 1 和表 2 所示。通过 PCK 指标对关键点检测准确性进行评估，当 T ＝0.1 时，CMU 数据集上，CSPDarkNet-VT 的 PCK 为 92.18%，FreiHAND 数据集上为 92.15%。CSPDarkNet-VT 在两个数据集上的 PCK 均明显高于除 CSPDarkNet＋Soft-Attention 以外的对比模型，且保持了与 CSPDarkNet＋Soft-Attention 相近的精度。从参数量和浮点运算来看，聚焦线性注意力机制有效减少了计算量，提高了计算效率。在精度方面，CSPDarkNet-VT 仍保持了与加入 Softmax 注意力相近的精度。在小目标检测任务中，聚焦线性注意力机制的表现更出色。这充分证明了 CSPDarkNet-VT 在小目标特征提取方

表 2　多个最先进模型作为特征提取器在 FreiHAND 数据集上的性能

Table 2　Performance of multiple state of the art models as feature extractors on the FreiHAND dataset

模型	参数量/M	浮点运算数/G	PCK/%
ResNet34	23.79	61.3	85.97
EfficientNetB1	8.42	10.1	86.83
EfficientNetB2	9.55	11.5	90.13
CSPDarkNet	3.39	9.7	88.09
CSPDarkNet＋Soft-Attention	3.26	9.4	92.19
CSPDarkNet-VT (Our)	3.04	8.7	92.15

注：PCK 为 T ＝0.1 时的值

面的强大能力。手部在图像中通常属于小目标范畴，准确提取手部的特征对姿态估计至关重要。CSPDarkNet-VT 能更有效地捕捉手部的细节特征，从而提高了关键点检测的准确率，为后续的姿态估计提供了更可靠的基础。

综上所述，CMU 和 FreiHAND 数据集上的对比实验表明，CSPDarkNet-VT 在参数量、计算效率和 PCK 指标上均表现出明显优势。因此，CSPDarkNet-VT 是一种更适合 2D 手部姿态估计任务的特征提取器，为提高姿态估计模型的性能提供了有力支持。

4.2 对比试验

不同误差阈值 T 下，本文在 CMU 的全景手数据集上和 FreiHAND 数据集上，将 FAR-HandNet 模型与其他几种姿态估计方法进行了比较，结果如表 3 和表 4 所示。FAR-HandNet 在两个数据集上的检测示例如图 4 所示。

在 CMU 数据集中，当 $T=0.04$ 时，CPM 的 PCK 为 55.25%，NSRM-LDM-G1 为 59.20%，NSRM-LPM-G1为59.81%，RetinaHand为60.12%，CH-HandNet为61.13%，FAR-HandNet为66.33%，FAR-HandNet 显著高于其他对比方法，表明其在低阈值下能更准确地检测手部关键点，对关键点位置的预测更精准。当 $T=0.12$ 时，CPM 的 PCK 为 88.80%，NSRM-LDM-G1 为 89.81%，NSRM-LPM-G1为90.26%，RetinaHand为90.74%，CH-HandNet为91.35%，FAR-HandNet为93.79%，FAR-HandNet 同样远超其他方法。在不同阈值下，FAR-HandNet 的平均 PCK 也表现出色，达到了 84.66%，超过了其他所有对比方法，表明 FAR-HandNet 在不同阈值下均能保持较高的检测准确率，具有较强的适应性和稳定性。在 FreiHAND 数据集中，当 $T=0.04$ 时，FAR-HandNet 的 PCK 同样高于其他模型；当 $T=0.12$ 时，FAR-HandNet 的 PCK 领先于其他方法，证明了其在不同数据集下的优秀性能。

通过对比分析不同阈值下的 PCK 和平均 PCK 可知，在手部姿态估计任务中，FAR-HandNet 能在不同精度下均保持较高的检测准确率。此外，在推理时间方面，FAR-HandNet 对单幅图像的推理时间仅需 32 ms，与其他对比方法相比，有了极大提升，这将为人机交互场景提供实时性更高的解决方案。

4.3 消融实验

本文在 CMU 的全景手数据集上进行消融实验，进一步检查本文模型每个模块的性能，如表 5 所示。在消融实验中，FAR-HandNet 的各模块均表现出色，表明其各模块在提高整体性能方面均发挥了重要作用。与原始模型 YOLOv8 相比，仅加入 RrFM 模块时，模型的平均 PCK 提升了 6.05%。RrFM 模块通过创新的回归范式，使得模型在处理手部姿态估计任务时，能更好地适应手部关键点的复杂分布，减少因传统回归方法的局限性而导致的误差，从而提升模型在关键

表 3 CMU 数据集上对比结果

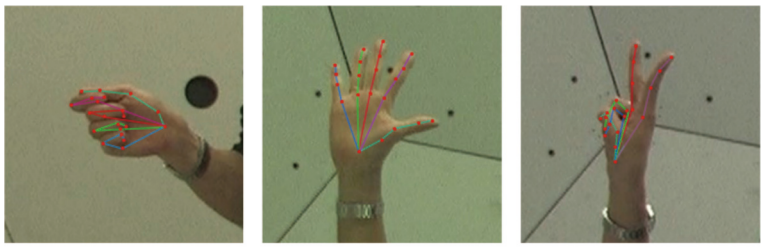
Table 3 Comparison results on the CMU dataset

方法	耗时/ms	PCK/%				$\overline{\text{PCK}}$/%
		$T=0.04$	$T=0.08$	$T=0.10$	$T=0.12$	
CPM[7]	131	55.25	81.45	86.73	88.80	78.06
NSRM-LDM-G1[23]	81	59.20	83.54	87.46	89.81	80.00
NSRM-LPM-G1[23]	83	59.81	84.16	87.96	90.26	80.55
RetinaHand[24]	56	60.12	83.63	88.11	90.74	80.65
CH-HandNet[25]	53	61.13	85.28	88.32	91.35	81.52
FAR-HandNet	32	66.33	86.34	92.18	93.79	84.66

表 4　FreiHAND 数据集上对比结果

Table 4　Comparison results on the FreiHAND dataset

方法	耗时/ms	PCK/%				$\overline{\text{PCK}}$/%
		$T=0.04$	$T=0.08$	$T=0.10$	$T=0.12$	
CPM[7]	131	54.86	81.15	86.48	88.56	77.76
NSRM-LDM-G1[23]	81	58.97	83.32	87.18	89.79	79.82
NSRM-LPM-G1[23]	83	59.63	84.03	87.69	90.15	80.38
RetinaHand[24]	56	59.92	83.49	87.97	90.54	80.48
CH-HandNet[25]	53	61.05	85.14	88.27	91.26	81.43
FAR-HandNet	32	66.28	86.29	92.15	93.74	84.62



(a) CMU 数据集

(b) FreiHAND 数据集

图 4　FAR-HandNet 在 CMU 和 FreiHAND 数据集上的姿态估计示例

Fig. 4　Example of pose estimation using FAR-HandNet on CMU and FreiRAND datasets

表 5　FAR-HandNet 的各模块性能分析

Table 5　Performance analysis of each module in FAR-HandNet

方法	耗时/ms	PCK/%				$\overline{\text{PCK}}$/%
		$T=0.04$	$T=0.08$	$T=0.10$	$T=0.12$	
YOLOv8-pose	64	58.59	79.37	85.77	86.31	77.51
RrFM	68	65.11	85.39	91.29	92.46	83.56
CSPDarkNet-VT	52	64.53	84.54	90.38	91.99	82.86
KPA	40	65.06	85.41	91.25	92.44	83.55
KPA＋CSPDarkNet-VT	28	65.12	85.39	91.28	92.48	83.59
FAR-HandNet	32	66.33	86.34	92.18	93.79	84.66

点检测上的准确性。当仅改进特征提取器为 CSPDarkNet-VT 时，平均 PCK 提高了 5.35%。CSPDarkNet-VT 在参数量和计算效率上具有明显优势，其独特的结构设计使其能更有效地提取小目标特征，更精准地捕捉手部的细节，为后续的姿态估计提供更丰富、准确的特征信息，进而提

升模型的整体性能。仅加入 KPA 模块时，平均 PCK 提升了 6.04%。KPA 模块通过对关键点空间关系和类别信息的综合考量，能更准确地定位关键点，提高模型在处理手部自相似问题时的能力，从而显著提升模型的精度。将 KPA 与 CSPDarkNet-VT 结合时，平均 PCK 提升了 6.08%。这进一步表明关键点对齐策略对密集的小目标区域更敏感，以及不同模块之间的协同作用能产生更显著的效果。CSPDarkNet-VT 提取的优质特征为 KPA 模块的关键点对齐操作提供了更好的基础，而 KPA 模块则充分利用这些特征，进一步优化了关键点的定位，两者的结合使模型在处理复杂小目标区域时的性能得到了大幅提升。FAR-HandNet 模型整合了这些模块的优势，平均 PCK 达到 84.66%，比原始模型有显著提升，充分展示了各模块组合使用时的增益效果。

由上述数据对比可知，RrFM、CSPDarkNet-VT 和 KPA 等模块在 FAR-HandNet 模型中都发挥着重要作用，它们各自从不同角度提升模型的性能，为实现高精度的 2D 手部姿态估计模型提供了有力支持。

综上所述，CSPDarkNet-VT 作为特征提取器在不同数据集上的表现优于其他模型，而 FAR-HandNet 在对比试验和消融实验中均显示出优异的性能和模块的重要性。这表明 FAR-HandNet 在两个数据集上的表现最优，能实现最佳的手部关键点检测效果。

5 结 论

本文提出的 FAR-HandNet 模型引入了聚焦线性注意力机制，采用多尺度特征融合策略扩大感受野，从而捕捉不同尺度的特征信息，增强小目标区域 (如手) 内的特征捕捉能力，且更注重手部细节。此外，通过关键点对齐操作，将关键点之间的空间关系与类别信息相结合，极大程度地降低手部自相似性对关键点定位准确性的影响，并加快了模型的训练。FAR-HandNet 还结合了一种新颖的回归范式 RrFM，通过流生成模型拟合残差分布，基于此获得每个关键点在图像中的概率分布，以回归的方式达到热图的效果，进一步提高手部姿态估计的精度。这些创新技术的应用使 FAR-HandNet 显著提升了手部姿态估计的精度，并降低了手自相似的影响，为手势识别和人机交互等应用提供了有力支持。

参 考 文 献

[1] Mehta D, Sridhar S, Sotnychenko O, et al. VNect: real-time 3D human pose estimation with a single RGB camera [J]. ACM Transactions on Graphics, 2017, 36(4): 1-14.

[2] Kulon D, Guler RA, Kokkinos I, et al. Weakly-supervised mesh-convolutional hand reconstruction in the wild [C] // Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 4990-5000.

[3] Bello I, Zoph B, Vaswani A, et al. Attention augmented convolutional networks [C] // Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, 2019: 3286-3295.

[4] Chen XJ, Yuille A. Articulated pose estimation by a graphical model with image dependent pairwise relations [J]. Advances in Neural Information Processing Systems, 2014: 27.

[5] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation [C] // Proceedings of the Computer Vision–ECCV 2016, 2016: 483-499.

[6] Wei SE, Ramakrishna V, Kanade T, et al. Convolutional pose machines [C] // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4724-4732.

[7] Wang YG, Peng C, Liu YB. Mask-pose cascaded CNN for 2D hand pose estimation from single color image [J]. IEEE Transactions on Circuits and

Systems for Video Technology, 2018, 29(11): 3258-3268.

[8] Wang YG, Zhang BW, Peng C. SRHandNet: real-time 2D hand pose estimation with simultaneous region localization [J]. IEEE Transactions on Image Processing, 2019, 29: 2977-2986.

[9] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks [C] // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1653-1660.

[10] Geng ZG, Sun K, Xiao B, et al. Bottom-up human pose estimation via disentangled keypoint regression [C] // Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 14676-14686.

[11] Carreira J, Agrawal P, Fragkiadaki K, et al. Human pose estimation with iterative error feedback [C] // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4733-4742.

[12] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale [Z/OL]. arXiv Preprint, arXiv: 2010.11929, 2020.

[13] Katharopoulos A, Vyas A, Pappas N, et al. Transformers are RNNs: fast autoregressive transformers with linear attention [C] // Proceedings of the International Conference on Machine Learning, 2020: 5156-5165.

[14] Han DC, Pan XR, Han YZ, et al. FLatten Transformer: vision Transformer using focused linear attention [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 5938-5948.

[15] Touvron H, Cord M, Douze M, et al. Training data-efficient image Transformers & distillation through attention [C] // Proceedings of the International Conference on Machine Learning, 2021: 10347-10357.

[16] Chollet F. Xception: deep learning with depthwise separable convolutions [C] // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251-1258.

[17] Redmon J, Farhadi A. YOLOv3: an incremental improvement [Z/OL]. arXiv Preprint, arXiv: 1804.02767, 2018.

[18] Ge Z, Liu ST, Wang F, et al. YOLOX: exceeding YOLO series in 2021 [Z/OL]. arXiv Preprint, arXiv: 2107.08430, 2021.

[19] Feng CJ, Zhong YJ, Gao Y, et al. TOOD: task-aligned one-stage object detection [C] // Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, 2021: 3490-3499.

[20] Li JF, Bian SY, Zeng AL, et al. Human pose regression with residual log-likelihood estimation [C] // Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, 2021: 11005-11014.

[21] Joo H, Simon T, Li XL, et al. Panoptic Studio: a massively multiview system for social interaction capture [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(1): 190-204.

[22] Zimmermann C, Ceylan D, Yang JM, et al. FreiHAND: a dataset for markerless capture of hand pose and shape from single RGB images [C] // Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, 2019: 813-822.

[23] Chen YF, Ma HY, Kong DY, et al. Nonparametric structure regularization machine for 2D hand pose estimation [C] // Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, 2020: 370-379.

[24] Xiao ZL, Lin LJ, Yang YX, et al. RetinaHand: towards accurate single-stage hand pose estimation [C] // Proceedings of the Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, 2023: 639-647.

[25] Zhang MY, Zhou ZH, Deng M. Cascaded hierarchical CNN for 2D hand pose estimation from a single color image [J]. Multimedia Tools and Applications, 2022, 81(18): 25745-25763.